Fully Bayesian analysis of RNAseq data for gene expression heterosis detection

Dr. Jarad Niemi

Iowa State University

July 30, 2018

Dr. Dan Nettleton (ISU) and Dr. Will Landau (Eli Lilly)

This research was supported by National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health and the joint National Science Foundation / NIGMS Mathematical Biology Program under award number R01GM109458. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

RNAseq data

Gene ID	B73				Mo17				B73 x Mo17				Mo17 x B73			
GRMZM2G107839	26	17	32	35	30	32	41	43	63	44	116	101	30	31	69	47
GRMZM5G899787	62	57	38	33	91	78	66	69	58	84	42	43	74	70	53	51
GRMZM5G899800	150	238	12	6	198	392	11	15	187	433	8	10	414	291	11	13
GRMZM2G301485	24	12	29	32	20	14	32	46	5	3	6	6	2	3	3	7
GRMZM5G899836	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

(Will Landau)

Low (high) parent heterosis: expression in hybrid is lower (higher) than both parents

Overdispersed count regression model

Let

- g ($g = 1, \ldots, G$) identify the gene,
- n ($n = 1, \ldots, N$) identify the sample,
- y be the $G \times N$ matrix of RNAseq counts and
- X be the $N \times L$ model matrix that connects the N samples to the varieties, blocking factors, etc.

We assume

$$y_{gn} \overset{ind}{\sim} \text{Po}\left(e^{h_n + x_n' \beta_g + \varepsilon_{gn}}\right)$$

where

- h_n are *normalization factors* (offsets),
- x_n is the n^{th} row of X,
- β_g is a vector of length L that account for effects on gene expression of variables of interest, and
- $\varepsilon_{gn} \overset{ind}{\sim} N(0, \gamma_g)$ allow for gene-specific overdispersion.

Hierarchical model

Recall

$$y_{gn} \overset{ind}{\sim} \text{Po}\left(e^{h_n+\varepsilon_{gn}+x_n'\beta_g}\right) \qquad \text{and} \qquad \varepsilon_{gn} \overset{ind}{\sim} N(0,\gamma_g).$$

We construct a hierarchical model for both β_g and γ_g to borrow information across genes. Specifically, we assume

$$\beta_{g\ell} \overset{ind}{\sim} N(\theta_\ell, \sigma_\ell^2)$$

for $\ell = 1, \ldots, L$ and

$$1/\gamma_g \overset{ind}{\sim} \text{Ga}\left(\nu/2, \nu\tau/2\right)$$

such that $E[1/\gamma_g] = 1/\tau$ and $CoV[1/\gamma_g] = \sqrt{2/\nu}$.

Heterosis hypotheses

Heterosis	With log-scale group means	With $\beta_{g\ell}$ parameters
high-parent BM	$\mu_{g,\text{BM}} > \max\left(\mu_{g,\text{B}}, \mu_{g,\text{M}}\right)$	$2\beta_{g2} + \beta_{g4},\quad 2\beta_{g3} + \beta_{g4} > 0$
low-parent BM	$\mu_{g,\text{BM}} < \min\left(\mu_{g,\text{B}}, \mu_{g,\text{M}}\right)$	$-2\beta_{g2} - \beta_{g4}, -2\beta_{g3} - \beta_{g4} > 0$
high-parent MB	$\mu_{g,\text{MB}} > \max\left(\mu_{g,\text{B}}, \mu_{g,\text{M}}\right)$	$2\beta_{g2} - \beta_{g4},\quad 2\beta_{g3} - \beta_{g4} > 0$
low-parent MB	$\mu_{g,\text{MB}} < \min\left(\mu_{g,\text{B}}, \mu_{g,\text{M}}\right)$	$-2\beta_{g2} + \beta_{g4},\quad 2\beta_{g3} + \beta_{g4} > 0$
high-parent mean	$\mu_{g,\text{BM}} + \mu_{g,\text{MB}} > 2\max\left(\mu_{g,\text{B}}, \mu_{g,\text{M}}\right)$	$\beta_{g2},\quad \beta_{g3} > 0$
low-parent mean	$\mu_{g,\text{BM}} + \mu_{g,\text{MB}} < 2\min\left(\mu_{g,\text{B}}, \mu_{g,\text{M}}\right)$	$-\beta_{g2}, -\beta_{g3} > 0$

All hypothesis regions are intersections of linear combination events, but we can also accommodate unions of contrast events via

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Priors

All priors are constructed to be vague, proper, and (if possible) conditionally conjugate. There are $2(L+1)$ hyperparameters and we assign the following priors

$$\begin{aligned}
\tau &\sim Ga(a, b) && \text{conditionally conjugate} \\
\nu &\sim Unif(0, d) && \\
\theta_\ell &\overset{ind}{\sim} N(0, c_\ell^2) && \text{conditionally conjugate} \\
\sigma_\ell &\overset{ind}{\sim} Unif(0, s_\ell) &&
\end{aligned}$$

As we'll see, posterior distributions for these parameters are extremely tight relative to their priors.

Constructing a Gibbs sampler

Conditional independence within a step:

$$
\begin{aligned}
p(\varepsilon|\ldots) &\propto \prod_{g=1}^{G}\prod_{n=1}^{N} Po\left(y_{gn}\middle|e^{h_n+\varepsilon_{gn}+x_n'\beta_g}\right) N(\varepsilon_{gn}|0,\gamma_g)\\
p(\gamma|\ldots) &\propto \prod_{g=1}^{G}\prod_{n=1}^{N} N(\varepsilon_{gn}|0,\gamma_g) IG(\gamma_g|\nu/2,\nu\tau/2)\\
p(\beta_\ell|\ldots) &\propto \prod_{g=1}^{G}\prod_{n=1}^{N} Po\left(y_{gn}\middle|e^{h_n+\varepsilon_{gn}+x_n'\beta_g}\right) N(\beta_{g\ell}|\theta_\ell,\sigma_\ell^2)
\end{aligned}
$$

Sufficient "statistics":

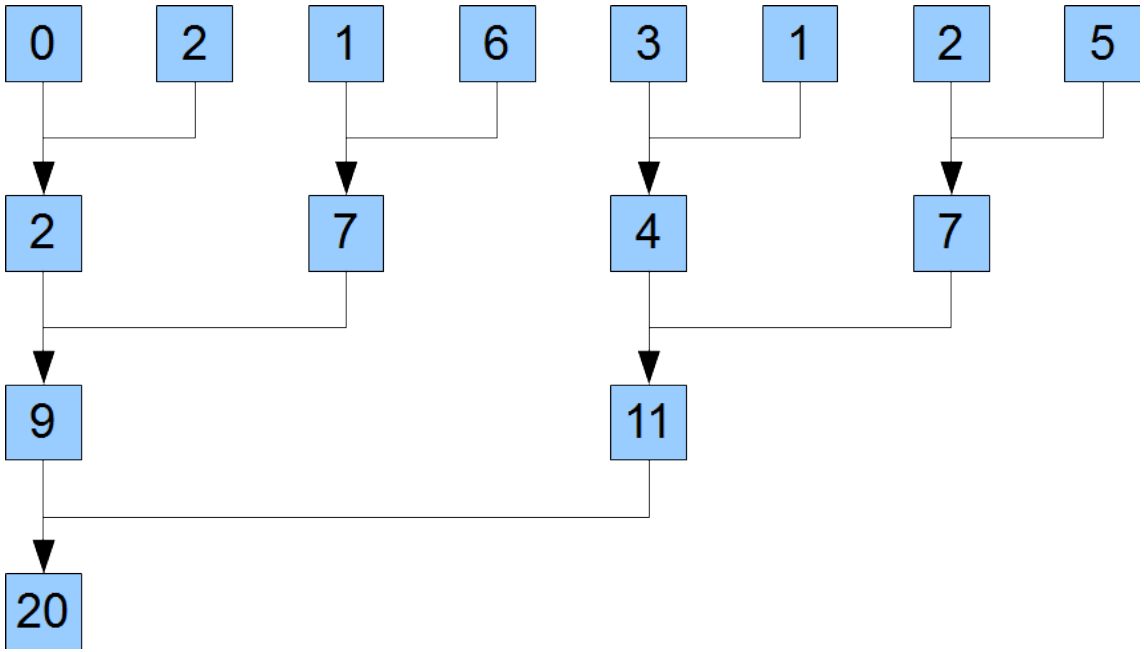
$$
\begin{aligned}
p(\tau|\ldots) &\sim Ga(\tau|a',b') & (a',b') &= f_\tau(\gamma,\nu,a,b)\\
p(\nu|\ldots) &\sim p(\nu|d')\mathrm{I}(0<\nu<d) & d' &= f_\nu(\gamma,\tau,d)\\
p(\theta_\ell|\ldots) &\sim N(\theta_\ell|m_\ell',C_\ell') & (m_\ell',C_\ell') &= f_{\theta_\ell}(\beta_\ell,\sigma_\ell,c_\ell^2)\\
p(\sigma_\ell^2|\ldots) &\sim IG(e',f')\mathrm{I}(0<\sigma_\ell^2<s_\ell^2) & (e',f') &= f_{\sigma_\ell}(\beta_\ell,\theta_\ell)
\end{aligned}
$$

where all functions can be written as sums over G terms in order to calculate means, variances, etc.

Parallelization translated to a GPU

If there are G nodes, then

- Conditional independence $\rightarrow$ embarrassingly parallel - possible speedup is G
- Calculate sufficient "statistics" $\rightarrow$ parallel reduction - possible speedup is $[G-1]/\log_2(G)$



(https://scs.senecac.on.ca/~gpu610/pages/images/parallel_reduction.png)

Implementation

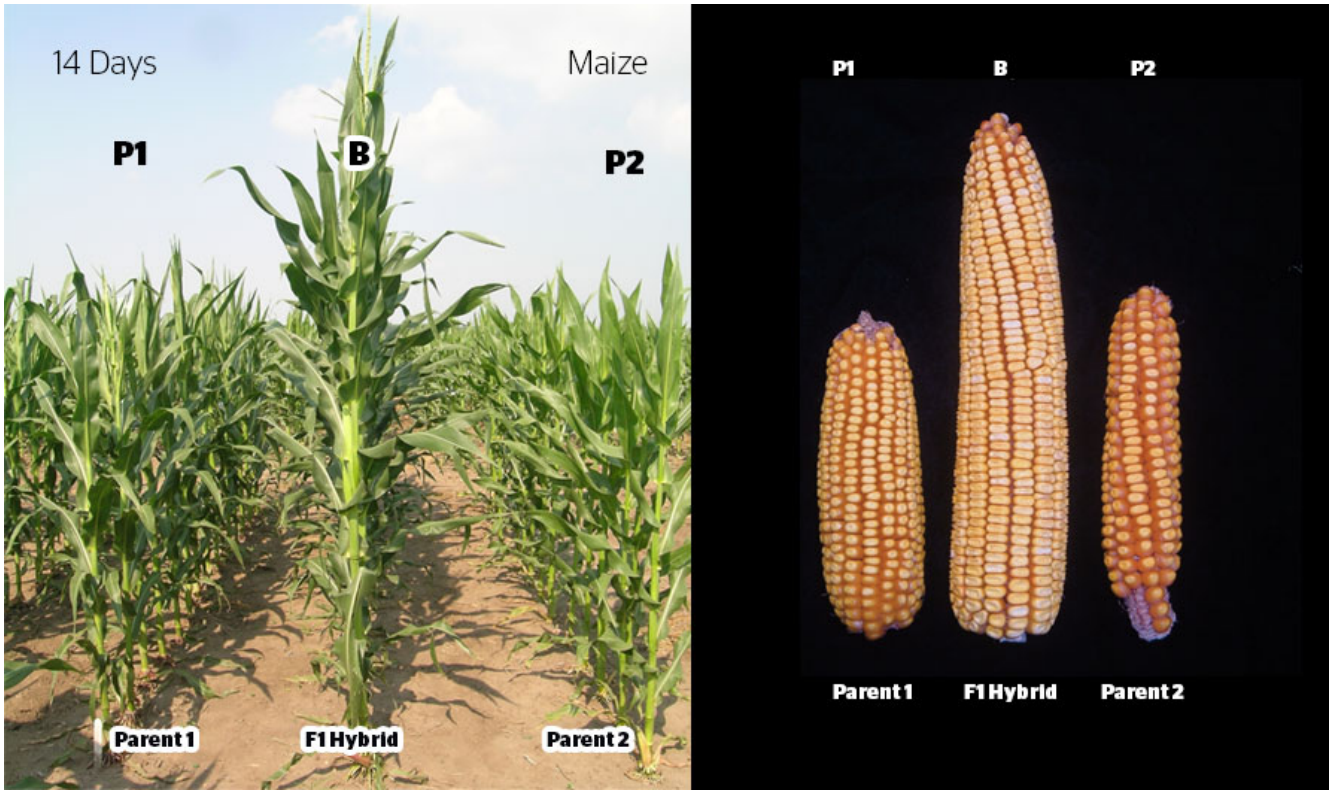
The computation for this hierarchical overdispersed count regression model is provided in the following three R packages at `https://github.com/wlandau/`:

- `fbseq`: user interface
- `fbseqOpenMP`: multithreaded backend
- `fbseqCUDA`: NVIDIA GPU backend

```
library(fbseq)
data(paschold) # Paschold et. al. (2012) data

paschold@contrasts[[5]]


## beta_1 beta_2 beta_3 beta_4 beta_5
##      0      2      0      1      0


paschold@contrasts[[6]]


## beta_1 beta_2 beta_3 beta_4 beta_5
##      0      0      2      1      0


paschold@propositions$`high-parent_B73xMo17`


## high-parent_B73xMo17_1 high-parent_B73xMo17_2
##                      5                      6
```

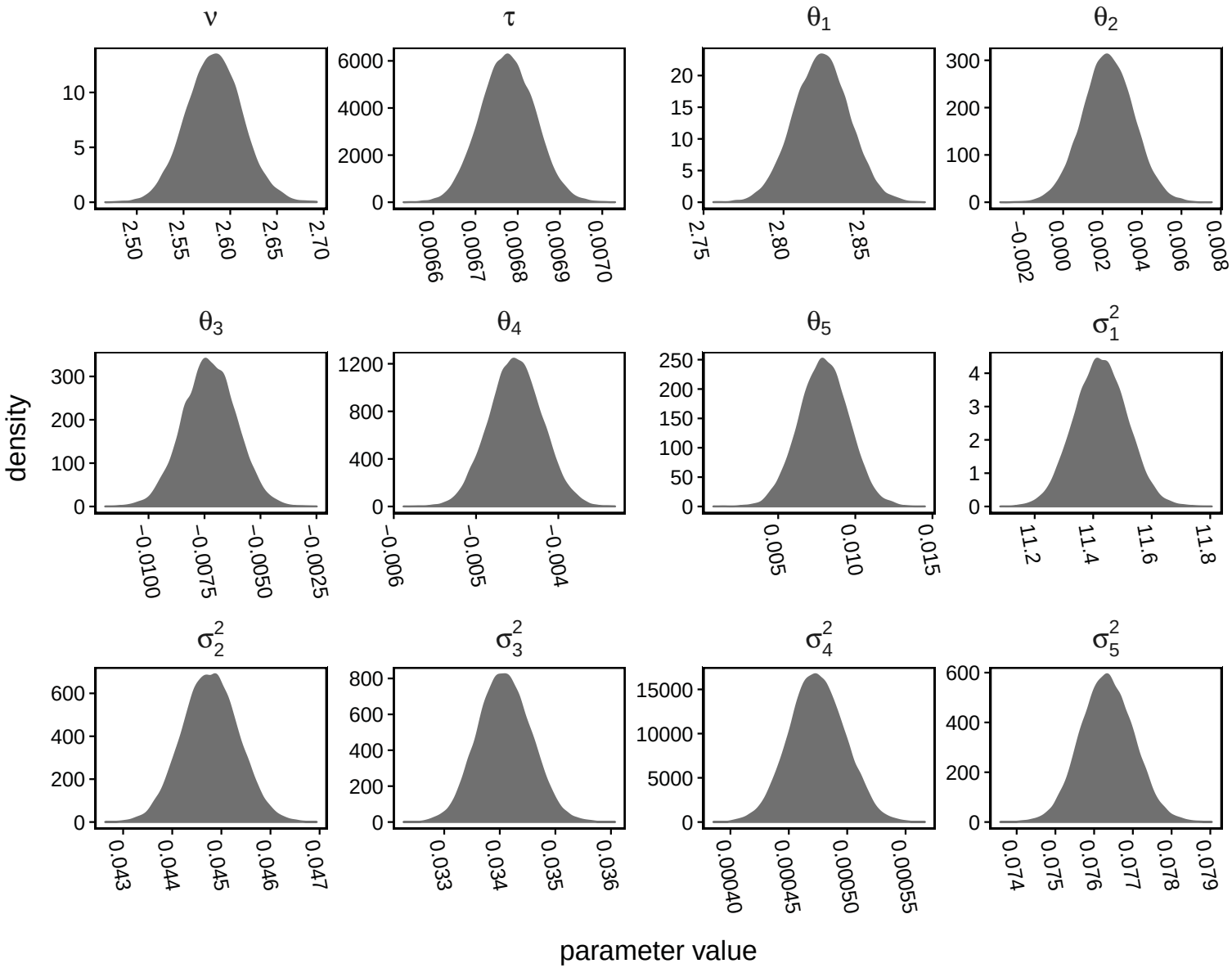
```
configs    = Configs(burnin = 10, iterations = 10, thin = 1)
chain      = Chain(paschold, configs)
chain_list = fbseq(chain, backend = "CUDA")
```

Analysis of Paschold et. al. (2012) data



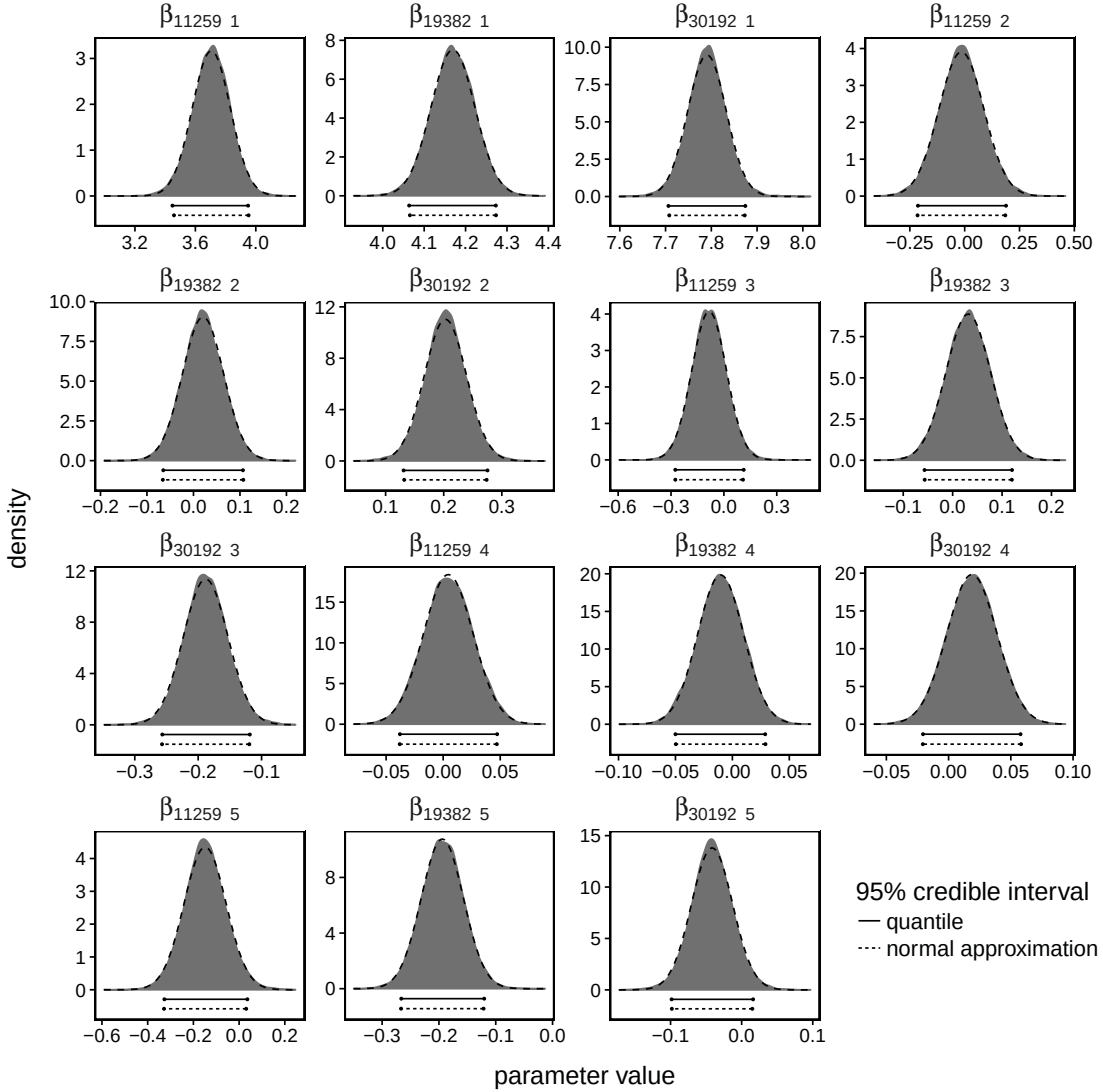
- $N = 16$ with 4 replicates/variety on 2 plates
  - varieties: B73, Mo17, B73$\times$Mo17, Mo17$\times$ B73
- $G = 39\,656$

Hyperparameter posterior

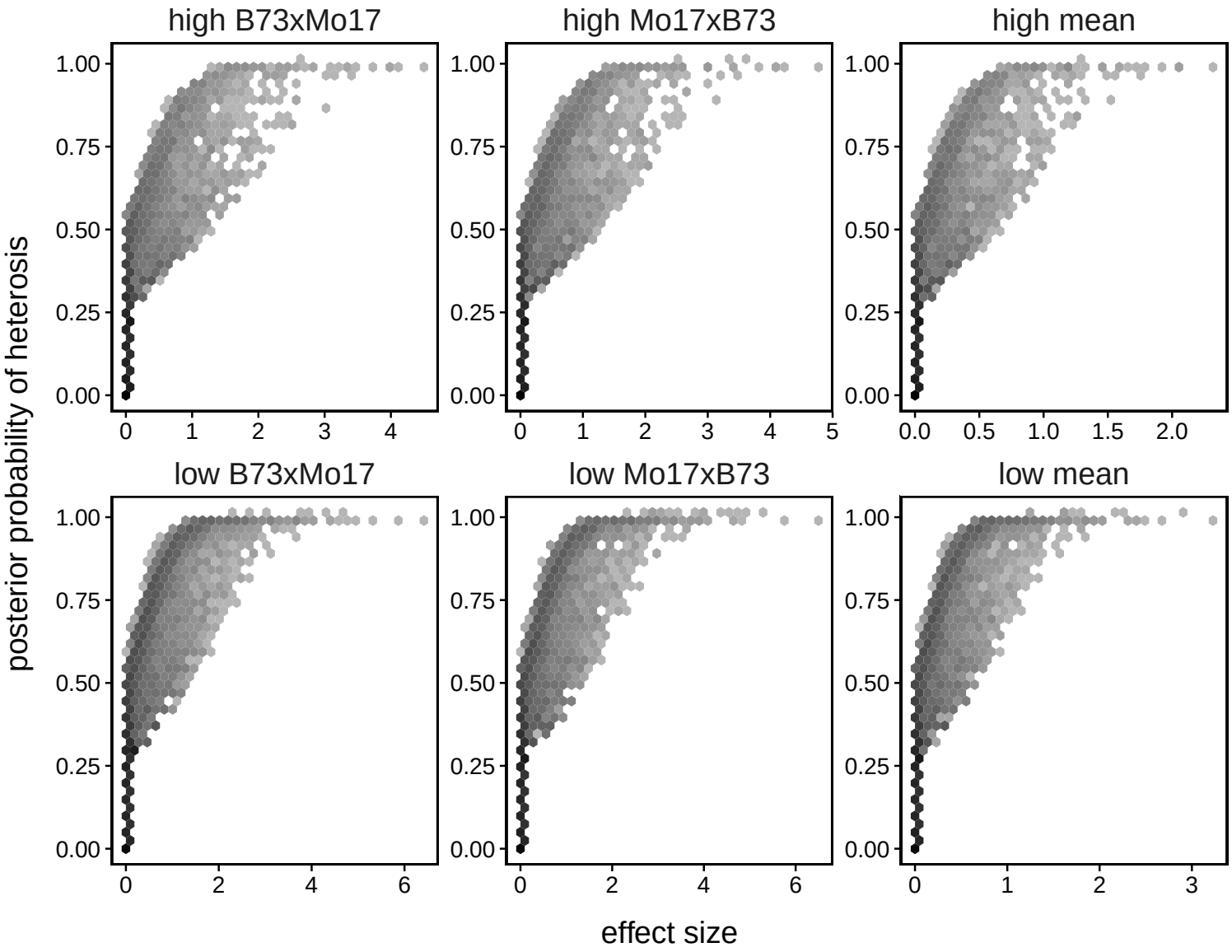


Gene-specific parameter posteriors

Compare posterior distribution (gray area) to normal-based approximation (black dashed line).



Smokestack plot



Summary

- Introduced a hierarchical overdispersed count regression model and
- a GPU-implementation of a fully Bayesian analysis.

This slides are available

- `https://github.com/jarad/JSM2018`
- `http://www.jarad.me/research/presentations.html`

Manuscripts:

- Jarad Niemi, Eric Mittman, Will Landau, and Dan Nettleton. (2015) "Empirical Bayes analysis of RNA-seq data for detection of gene expression heterosis." Journal of Agricultural, Biological, and Environmental Statistcs, 20(4): 614-628
- Will Landau and Jarad Niemi. "A fully Bayesian strategy for high-dimensional hierarchical modeling using massively parallel computing." `https://arxiv.org/abs/1606.06659`
- Will Landau, Jarad Niemi, and Dan Nettleton. "Fully Bayesian analysis of RNA-seq counts for the detection of gene expression heterosis." *Journal of the American Statistical Association* (in press)

Thank you!